Ontologia 2203NFC-*e*: Sentença completa do produto cerveja no campo descrição do produto da NFC-*e*

Diana Maria Camara Gorayeb,[1,†] *and* Claudio Gottschalg Duque[1,†]

[1] *Universidade de Brasilia, 70910-900 , Brasília, DF, Brasil*

Abstract

In this study called "2203NFC-e", the objective is to develop an ontology that will present the various versions of the "complete sentence" to replace the product description visible in the NFC-e field. From this process there will be a system for generating feedback results to improve user queries, indicating which NFC-e transactions may or may not be selected and used for various state tax purposes. For the construction methodology of the 2203NFC-e ontology, a simplification of ON-ODM was used, incorporating the steps of survey, analysis, conceptualization, implementation, enrichment, validation of the case study through Competency Questions (QC), formulated to meet inspection demands and overcome the problems presented. As a result, the 2203NFC-e ontology presents the reproduction of the beer product by complete sentence with description of the name of the beer, type of packaging, volume, package, quantity and individual value of any type of disordered description of the product, evaluating which are complete and which are useful in terms of mining one or more properties of the main term, allowing progress in the reading and interpretation of beer sales data for inspection.

Keywords

Information Science; 2203NFC-e ontology; beer product; complete sentence.

Resumo

Neste estudo denominado "Ontologia 2203NFC-e", objetiva-se elaborar uma ontologia que apresentará as várias versões da "sentença completa" para substituir a descrição do produto visível no campo da NFC-e. A partir desse processo haverá um sistema de geração de resultados retroalimentado para o aprimoramento das consultas dos usuários, indicando quais transações da NFC-e poderão ou não ser selecionadas e utilizadas em diversas finalidades do fisco estadual. Para a metodologia de construção da ontologia 2203NFC-e foi utilizada uma simplificação, da ON-ODM incorporando as etapas de: levantamento, análise, conceitualização, implementação, enriquecimento, validação do estudo de caso por meio das Questões de Competência (QC), formuladas para atender a demanda da fiscalização e superar os problemas apresentados. Como resultado, a ontologia 2203NFC-e apresenta a reprodução do produto cerveja por sentença completa com descrição do nome da cerveja, tipo da embalagem, volume, pacote, quantidade e valor individual de qualquer tipo de descrição desordenada do produto, avaliando quais são completas e quais são úteis em termos de garimpo de uma ou mais de uma propriedade do termo principal permitindo um avanço na leitura e interpretação dos dados de venda de cerveja para a fiscalização.

Palavras-chave

Ciência da Informação; ontologia 2203NFC-*e;* produto cerveja; sentença completa.

1. Introdução

A identificação das mercadorias comercializadas com Nota Fiscal de Consumidor Eletrônica (NFC-*e*) é feita por meio do seu correspondente, o código estabelecido na Nomenclatura Comum do Mercosul (NCM) [1]. O campo do NCM acompanha a descrição das mercadorias comercializadas em um formulário com: numeração do item, descrição do item (campo livre), quantidade, unidade comercial e valor. A dificuldade para o trabalho de análise tributária e acompanhamento do Fisco está neste

Proceedings of the 17th Seminar on Ontology Research in Brazil (ONTOBRAS 2024) and 8th Doctoral and Masters Consortium on Ontologies (WTDO 2024), Vitória, Brazil, October 07-10, 2024.

[†] These authors contributed equally.

diana.gorayeb@aluno.unb.br (D. Gorayeb); klauss@unb.br (C. Duque)

0009-0006-5081-4485 (D. Gorayeb); 0000-0003-3558-466X (C. Duque).

campo livre, pois não há padronização para representar o item comercializado nem controle entre o NCM informado e a descrição do item, o que dificulta a fiscalização utilizando as informações do documento eletrônico uma vez que cada emitente da NFC-*e* descreve o produto de uma forma diferente e quase sempre de forma incompleta, sem caracterizá-lo corretamente, quer seja por não existir formalmente nas legislações termos e padrões para esta tarefa, quer seja por uma intensão maliciosa para enganar o fisco.

Este estudo propõe construir uma ontologia denominada 2203NFC-*e* para apresentar as várias versões de uma "sentença completa", para substituir a descrição do produto visível no campo da NFC-*e*. A partir desse processo, haverá um sistema de geração de resultados retroalimentado para o aprimoramento das consultas dos usuários, indicando quais transações da NFC-*e* poderão ou não ser selecionadas e utilizadas em diversas finalidades do fisco estadual como: quais fornecedores não adotam uma descrição do produto de acordo com as exigências da legislação, composição do Preço Médio Ponderado ao Consumidor Final (PMPF), inconsistência entre descrição do produto e do código de produto para fins de pagamento de impostos etc.

Para melhor organização, este trabalho seguiu 4 etapas, quais sejam: a base teórica sobre ontologia, resumo de alguns trabalhos relacionados à temática abordada, metodologia e desenvolvimento da ontologia. Essas etapas possibilitaram chegar a um resultado com relevância no campo da Ciência da Informação e na área fiscal.

2. Ontologia

Sistemas de gestão baseados em conhecimento utilizam ontologia para dar suporte aos processos organizacionais usando ferramentas automáticas. Assim sendo, a ontologia descreve explicitamente uma conceitualização compartilhada, uma interpretação estruturada de uma parte do mundo que as pessoas usam para pensar e se comunicar [5]. Além disso, as ontologias fornecem contexto e significado aos dados e são fundamentais na extração e reutilização do conhecimento, proporcionam uma solução robusta para o problema da interoperabilidade sintática e semântica que dificulta a troca de informações em sistemas heterogêneos [6].

De modo sintético, ontologias são compostas de "conceito", um modelo; uma expressão humana do mundo real, parecido ao significado de "classe" na orientação a objetos; um "relacionamento", um conceito entre conceitos ou uma associação de classes; "instância", o elemento básico do conceito ou um exemplo concreto; "função", uma descrição abstrata do método; "axioma", um fato reconhecido ou regra de inferência [12].

Existem várias metodologias de desenvolvimento de ontologias, há de se considerar aquela que é mais adequada a partir do objetivo, das características do domínio e da fonte de conhecimento, como destaca [6]. Inúmeras metodologias apresentam diferentes perspectivas do processo e do foco combinando as fases e expandindo-as com novas ferramentas, inclusive de gestão e de Inteligência Artificial (IA). Foram encontrados, em fontes de pesquisa acadêmicas, estudos com metodologias relevantes e aprimoradas utilizando instrumentos de IA como: redes neurais, *clusters*, regras de associação, PLN, entre outros.

Um resumo da metodologia adaptada à este trabalho: ON-ODM – *Ontology Development Methodology,* [6]: 1º. Aquisição dos Requisitos: identificação do domínio, análise e especificação em Questão de Competência (QC); 2º. Conceitualização: especificação dos conceitos, definição em diagrama de classes da UML, formalização em linguagem lógica OWL, implementação com PROTÉGÉ; 3º. Enriquecimento por meio de PLN: **a.** Utilização de um Corpus texto para candidatos a novos termos; **b.** Segmentação de frases e busca das classes; **c.** *Tokenização* das frases e atribuição de tipos sintáticos aos tokens; **d.** Extração dos verbos para futuros relacionamentos entre classes; **e.** Lista final de candidatos para enriquecer a ontologia; 4º. Avaliação: verificação baseada em questões de competência e em métricas (precisão, coesão, compreensão, concisão); 5º. Publicação; 6º. Manutenção; 7º. Documentação.

Uma adaptação do planejamento para elaboração do Sistema de Organização do Conhecimento (SOC) no domínio NFC-*e*, apresentado por [3], incluiu as etapas do Modelo IA e de Ontologias

Referenciadas para extrair conjunto dos termos e atributos classificados e as associações mais significativas em um processo semiautomatizado, sem a interferência de especialistas, extraindo, diretamente da base de dados NFC-*e,* termos que se repetem e tem relevância para o processo de descrição do produto. O resumo do planejamento, dos processos e artefatos gerados para a NFC-*e* é apresentado a seguir:

1. Levantamento e compreensão das normas e documentos da gestão do fisco (Resolução n.º 0028/2023 SEFAZ/AM) e definição do projeto para fiscalização:
    a. Filtros para segmentos mais importantes para arrecadação;
    b. Lista de produtos de substituição tributária com PMPF;
    c. Nome dos produtos que poderão enriquecer a ontologia.
2. Entrada dos filtros que serão aplicados na base Nota Fiscal Eletrônica (NF-*e*) e NFC-e:
    a. NCM;
    b. Indicação de termo principal;
    c. Períodos pré-determinados de arrecadação etc.
3. Definição dos metadados para descrição dos produtos:
    Sentença completa e útil à fiscalização.
4. Levantamento e definição de características sintáticas e semânticas dos termos, atributos e associações por meio de algoritmos de IA:
    a. Ontologia: lista de candidatos às classes, subclasses, qualificadores (atributos) e relações existentes
5. Definição, validação e enriquecimento dos termos:
    a. Reuso de ontologias referenciadas; e
    b. utilização de termos adicionais que estão nas Resoluções da SEFAZ/AM.
6. Incorporar propriedade de dados:
    a. Colocar os dados das NFC-e como *Data Property Assertions* e instâncias das classes da ontologia para validação dos dados.

A implementação do processo 3: Definição dos metadados para descrição dos produtos propôs termos em quantidade e qualidade suficientes para descrever o produto cerveja apresentado no campo “descrição do produto” a partir da mineração de dados a base da NFC-*e* usando Apriori [4]. O estudo apresenta um procedimento metodológico com sete etapas para extração e categorização de termos frequentes para propor metadados de descrição do produto utilizando um filtro de frequência mínima do termo (800 repetições mínimas). Das associações de 3, 4 e 5 termos o estudo pôde afirmar quais termos representavam com mais convicção os atributos da descrição da cerveja, oferecendo uma ordem apresentação: “nome do produto + nome da marca do produto + tipo de embalagem do produto + capacidade da embalagem do produto”. O resultado deste estudo será aplicado para construção e apresentação da sentença completa neste trabalho.

3. Trabalhos Relacionados

Para ampliar o conhecimento acerca da temática, buscou-se ontologias que foram criadas utilizando metodologias diversas e com aproximação ao conteúdo deste trabalho. Assim, seis trabalhos que contribuem para este estudo são descritos abaixo:

Alguns trabalhos utilizam técnicas de construção de ontologias juntamente com aprendizado não supervisionado e PLN. Um estudo utilizando Apriori e Ontologia [2] a partir de palavras-chave médicas, definidas pelos especialistas em uma base de dados relativamente limitada e explícita, permitiu a construção de uma ontologia a partir do significado e da associação dos atributos extraídos por mineração de texto. Neste caso, foi usando um mecanismo de inferência para inferir a relação de associação entre palavras médicas, doenças e complicações, criando uma base de conhecimento.

Outra ontologia estudada foi de risco financeiro logístico [12]. As palavras-chave foram levantadas com especialistas da área de risco e uma conceitualização e classificação formal foi

construída. A partir disso, o algoritmo de mineração de regras de associação Apriori foi aplicado para inferir regras entre o risco, evento de risco e controles.

Em [6] uma ontologia realizada no campo do turismo, com ajuda de especialistas, para enriquecer novos relacionamentos entre as classes. Assim sendo, alguns textos de interesse turísticos foram minerados com as técnicas de PLN: extração de *tokens*, POS, *Lemmatization* para identificação de verbos que pudessem ser utilizados para identificação de relacionamentos, conforme avaliação dos especialistas.

Especificamente sobre ontologias de cerveja [10] recorreu-se a uma ontologia de tipos de cerveja, a qual trata a discrepâncias na descrição e nos rótulos das cervejas e oferece recomendações de cerveja a partir de preferências de teor alcoólico, amargor, doçura, cor e ingredientes fornecidos por especialistas em cervejas.

Seguindo no campo da cerveja, *The Beer Ontology* [11] também apresenta uma ontologia de cervejas com estilos, legislação, fabricação, recipiente de fabricação e recipientes de embalagem das cervejas, resultando em um inventário para ajudar na tomada de decisão para seleção de cervejas. Essa ontologia definiu uma superclasse para *Packing* que estudou e definiu classes de tipos de embalagens como 'barril', 'garrafa', 'lata' e instâncias com suas volumetrias e empacotamentos: '12 garrafas de 355ml' ou '12 latas de 355ml'. Essas descrições são encontradas em muitas transações das NFC-*e* analisadas neste trabalho.

Por fim, analisou-se um trabalho que trata de uma ontologia de compra e venda de produtos [9]. Acompanha o processo do pedido de compra do produto, emissão de nota fiscal, envio do produto, recepção e pagamento. Sua relevância para este trabalho é a conceitualização dos campos da nota fiscal como item do produto, preço, códigos, quantidade, descrição etc. e que serão formalizadas como classes na ontologia. O trabalho também apresenta uma lista de questões de competência significativas para o contexto da fiscalização como: qual número da nota fiscal; quais itens são listados na nota fiscal; qual quantidade e preço dos itens; e quais atributos do item produto.

Neste trabalho, como já mencionado na seção 1, será construída uma ontologia que descreve a comercialização do produto cerveja, NCM 2203.xxxx, que abstrai as características relevantes à venda da cerveja como nome e marca, quantidade do volume comercializado, embalagem, preço, data, e demais informações sobre emissor e consumidor. As sentenças completas que descrevem a transação de venda do produto serão validadas por meio das QCs escolhidas a partir dos requisitos necessários à fiscalização.

4. Metodologia para o desenvolvimento do projeto

Esta é uma pesquisa com fins práticos com utilização e consequência prática dos conhecimentos [14], o método propõe o modelo de identificação do problema e possível solução [15] com objetivo de pesquisa exploratória com análise qualitativa [16] para avaliar trabalhos relacionados com a área de ontologia e IA. A coleta de dados inclui amostra fornecida pela SEFAZ/AM. Para a metodologia de construção da ontologia 2203NFC-*e* foi utilizada uma simplificação da ON-ODM, a escolha foi baseada na divisão de etapas necessárias para construção da ontologia propostas pela metodologia, um passo a passo bem definido, extenso, porém flexível que permitiu incorporar nos elementos de computação o módulo de IA proposto em [3]. As etapas selecionadas são: 1. levantamento; 2. análise; 3. conceitualização; 4. implementação; 5. enriquecimento; 6. ferramentas de Tecnologia da Informação e Comunicação (TIC) bem definidas; 7. validação do estudo de caso por meio das Questões de Competência (QC): QC01: Quais os termos que compõem uma expressão significativa para descrição do produto, ou seja, qual sentença completa pode ser utilizada para descrever corretamente a venda da cerveja de malte? QC02: é possível identificar descrições de produto em desacordo com o código de NCM em uma NFC-*e*? QC03: é possível identificar qual a quantidade e preço de determinado produto em determinado período? QC04: é possível identificar quais descrições de produto são significativas ou não significativas para um determinado uso?

As ferramentas utilizadas são: Knime, para aplicação dos algoritmos de AM; software Protegé para a implementação da ontologia 2203NFC-*e* com a linguagem OWL e sintaxe *Turtle,* consultas em DL *Query* e os *plugins*: DL *Query* 4.0.1, *Hermit Reasoner* 1.4.3 / 456, OntoGraf 2.0.3 e OWLViz 5.0.3.

5. Desenvolvimento da Ontologia 2203NFC-*e:* leitura e interpretação dos dados

Os dados para esse estudo foram disponibilizados pela Secretaria de Fazenda do Estado do Amazonas (SEFAZ/AM) em arquivo .csv, tipo texto, no período de 01/02/2023 a 31/05/2023: 2203.xxx – Cerveja de malte. Uma vez conhecida e disponível a base de dados, foi escolhido o NCM com os 4 dígitos iniciais 2203.xxxx: "Cerveja de malte" para a construção do modelo. A seleção da amostra relevante para NCM 2203.xxxx apresentou 4.019.340 transações.

O algoritmo Apriori foi utilizado para investigar a existência de termos frequentes na forma proposta em [4] entretanto, os filtros que os autores estabeleceram não foram aplicados neste estudo para não restringir a base de dados e permitir investigar todas as descrições dos produtos das NFC-*e* e novos termos se tornaram potencialmente candidatos às classes e instâncias da ontologia. Os resultados apresentam os termos de maior Suporte: 0,402 para "CERVEJA" e de 0,357 para "CERV", que são significativos para o item de interesse "Cerveja de malte", NCM: 2203.xxxx.

Ao final, da análise da frequência de ocorrência das palavras 210 termos são propostos pelas regras do Apriori para a definição dos conceitos da ontologia da venda do produto cerveja.

ITEM	SUPORTE	TERMO	ITEM	SUPORTE	TERMO	ITEM	SUPORTE	TERMO	ITEM	SUPORTE	TERMO	ITEM	SUPORTE	TERMO
1	0,000501406	sixpack	43	0,000812028	12x1l	85	0,001582438	gel	127	0,003449425	und	169	0,012450170	tijuc
2	0,000503663	caipirinh	44	0,000815036	lima	86	0,001591464	trig	128	0,003524886	ipa	170	0,012762295	extra
3	0,000504665	tropical	45	0,000829075	appi	87	0,001594723	witbi	129	0,003525388	pm	171	0,014479361	teor
4	0,000513440	lour	46	0,000831332	congel	88	0,001609264	kg	130	0,003536669	210ml	172	0,017662539	cor
5	0,000523468	pi	47	0,000831332	maring	89	0,001622802	caix	131	0,003901443	cervitaip	173	0,018593651	crystal
6	0,000530989	24x330ml	48	0,000838853	weiss	90	0,001632830	c15	132	0,003940051	chop	174	0,022429910	arto
7	0,000535251	pma	49	0,000850134	unic	91	0,001645616	1x355ml	133	0,003962363	sh	175	0,022487071	355ml
8	0,000537006	prom	50	0,000850887	amber	92	0,001685477	npal	134	0,004219836	c12	176	0,022875410	300ml
9	0,000540516	350mlcr	51	0,000858408	artesanal	93	0,001710046	15x269ml	135	0,004532462	beats	177	0,025767772	stell
10	0,000547034	sub	52	0,000875205	heinek	94	0,001715311	vd	136	0,004549009	happy	178	0,028614507	600ml
11	0,000558817	35l	53	0,000895512	ice	95	0,001722331	473ml	137	0,004571321	dobr	179	0,029256056	sleek
12	0,000559068	1x300ml	54	0,000899022	atac	96	0,001723334	cristal	138	0,004721242	litra	180	0,030896156	original
13	0,000584389	12x269	55	0,000937379	golden	97	0,001741886	longneck	139	0,004745811	becks	181	0,032681413	lag
14	0,000587398	cx15	56	0,000937881	cervskol	98	0,001759685	eisen	140	0,004877931	baden	182	0,035054068	amstel
15	0,000587648	990ml	57	0,000939886	pack	99	0,001765201	brasil	141	0,004901247	1l	183	0,038580960	neck
16	0,000593164	nec	58	0,000942142	330355l	100	0,001769964	pmalt	142	0,005057435	orig	184	0,038988353	antarct
17	0,000593665	cervbrahm	59	0,000968717	negr	101	0,001781998	garraf	143	0,005083759	subzer	185	0,042131418	bohem
18	0,000597927	frut	60	0,000975235	6x330ml	102	0,001789268	cabar	144	0,005126629	lt269ml	186	0,042251254	spaten
19	0,000609710	fant	61	0,000976740	malzbi	103	0,001805313	lon	145	0,005162479	one	187	0,045696166	imperi
20	0,000615226	ext	62	0,000987770	sol	104	0,001806316	ambev	146	0,005283569	gf	188	0,047557888	dupl
21	0,000622245	unidad	63	0,000993537	gross	105	0,001809826	imper	147	0,005390118	premium	189	0,048736193	long
22	0,000634028	embtest	64	0,001000055	dup	106	0,001854451	patagon	148	0,005647089	way	190	0,050147652	budweis
23	0,000634530	escur	65	0,001001308	art	107	0,001876764	prem	149	0,005706255	proib	191	0,060514478	lta
24	0,000635783	beb	66	0,001022117	guar	108	0,001934175	sle	150	0,006030665	cerp	192	0,069857683	un
25	0,000647566	spat	67	0,001025877	12un	109	0,001936431	cx-12	151	0,006081808	12x350ml	193	0,073610459	skol
26	0,000650575	gin	68	0,001033900	camar	110	0,002064541	coronit	152	0,006269334	petr	194	0,078677671	chopp
27	0,000659349	343ml	69	0,001056714	rio	111	0,002089360	litr	153	0,006310449	lt350ml	195	0,079173061	ml
28	0,000661856	bud	70	0,001060474	ale	112	0,002143512	lneck	154	0,006656670	12x269ml	196	0,080636666	kais
29	0,000668124	pc12	71	0,001116883	unfiltered	113	0,002302959	gfa	155	0,006689011	1x330ml	197	0,081506856	pur
30	0,000669628	munichpur	72	0,001135435	350g	114	0,002333545	cer	156	0,006939965	ow	198	0,083331725	lat
31	0,000669879	cv	73	0,001141201	hour	115	0,002403993	gold	157	0,007783330	tig	199	0,102474415	itaip
32	0,000676899	5l	74	0,001203375	fi	116	0,002469175	glacial	158	0,007857037	devass	200	0,103036492	ln
33	0,000676899	1lt	75	0,001222679	pil	117	0,002523077	divers	159	0,008189971	la	201	0,110538032	heineken
34	0,000690938	gluten	76	0,001249254	ma	118	0,002527088	beer	160	0,008651265	munich	202	0,118773380	330ml
35	0,000705980	boh	77	0,001287862	caracu	119	0,002529846	pils	161	0,008775864	schin	203	0,123554289	brahm
36	0,000709991	pal	78	0,001325718	grf	120	0,002588009	350m	162	0,009254456	can	204	0,149042026	pilsen
37	0,000726287	laranj	79	0,001328225	export	121	0,002628121	pr	163	0,009357245	cx	205	0,158234809	malt
38	0,000758377	stel	80	0,001345524	ita	122	0,002660211	sens	164	0,010076512	eisenbahn	206	0,203171345	269ml
39	0,000771664	12x350	81	0,001358310	1x600ml	123	0,002789574	color	165	0,010735611	500ml	207	0,255103752	lt
40	0,000772166	dmalt	82	0,001380622	dm	124	0,002793335	250ml	166	0,010908094	1x350ml	208	0,272461186	350ml
41	0,000793476	cozumel	83	0,001462101	gt	125	0,003026489	269m	167	0,011999656	antart	209	0,369717214	cerv
42	0,000794478	american	84	0,001463354	bra	126	0,003218527	275ml	168	0,012039518	1x269ml	210	0,413370902	cervej

Figura 1: Lista de 210 termos frequentes obtidos com a aplicação do Apriori. Fonte: Dados da pesquisa, 2024

O Apriori foi executado novamente para apresentar a associação entre 2, 3 e 4 termos do *itemset* e analisar a força das regras entre os termos de maior suporte do Corpus mantendo Suporte: 0,0001 e definindo a Confiança em 0,9. O objetivo é extrair candidatos para *Object Property*, relações entre as classes, e eliminar as relações fora do contexto de interesse. A seguir, a sequência de atividades foi realizada para construção da ontologia 2203NFC-*e*:

- Seleção do termo principal: o estudo demonstra que, independentemente do termo de interesse “cerveja” estar na descrição do NCM 2203.xxxx como Cerveja de malte, ele também se repete na forma “CERV” e “CERVEJ”;
- Seleção das classes a partir estudo [4];
- Seleção dos metadados da Nota Fiscal correspondentes ao produto: item da nota;
- Seleção dos metadados da Nota fiscal, correspondente à venda do produto: Unidade da Federação (UF), Município, dados do Emitente (Grupo C da NF-*e*), dados do Destinatário (Grupo E da NF-*e*) Número da Nota Fiscal ou Cupom Fiscal, Descrição do produto e serviço (Grupo I da NF-*e*);
- Seleção dos significados: alguns termos relacionados ao produto cerveja foram definidos com Michaelis Dicionário Brasileiro da Língua Portuguesa[2], [11]; [1] e [8];
- Enriquecimentos: com os trabalhos [11]; [9]; e [8];
- Definição dos metadados para construção da “sentença completa”: foram escolhidos o termo principal associado ao nome do produto + nome da marca do produto + embalagem (tipo e volume) propostos em [4] + pacote + NCM + cEAN (código GTIN) + número do item + quantidade + unidade + valor + descrição original (extraído do campo “descrição do produto” da nota fiscal).
- Adicionar a descrição original como instância para que a fiscalização possa comparar o resultado da sentença completa com a sentença original

Na Figura 2 abaixo, a ontologia 2203NFC-*e* implementada no Protégé:

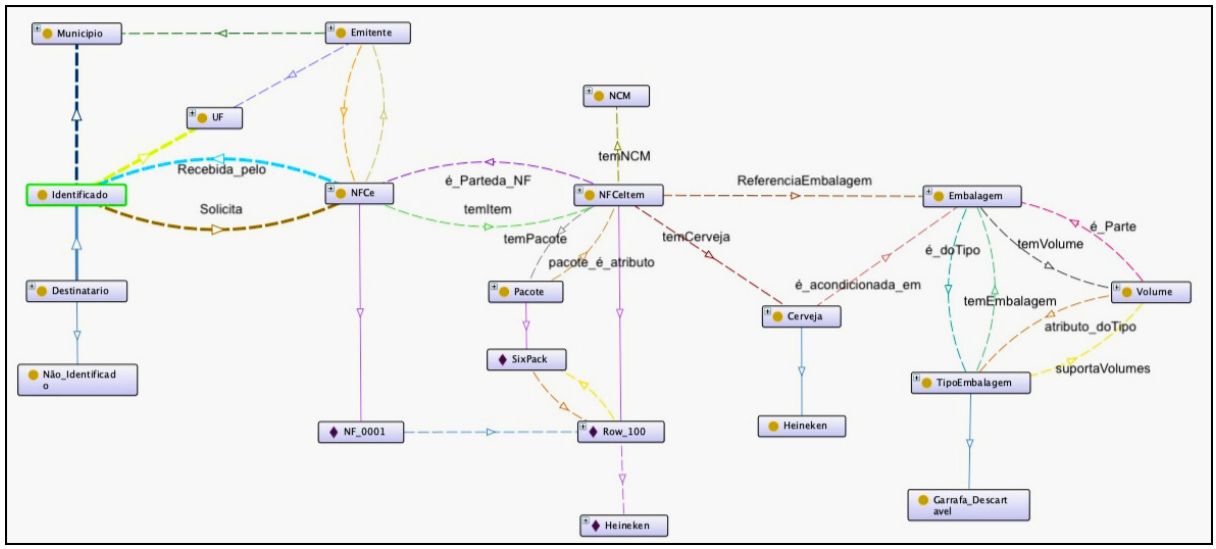

Figura 2: Destaque das classes e relacionamentos da ontologia 2203NFC-e [13]. Fonte: Dados da pesquisa, 2024

5.1. Questões de Competência

As respostas das QCs foram extraídas da ontologia 2203NFC-*e* e apresentadas para validar a apresentação da sentença completa de uma NFC-*e*, também as sentenças semicompletas que poderiam ser utilizadas em consultas especificas para a fiscalização ou para reconhecer sentenças erradas de descrição do produto. A extração foi feita utilizando uma linguagem de consultas de dados (DL *Queries*) disponível no Protégé.

[2] Dicionário online, acesso em fevereiro/2024.

Na análise da QC01, qual a sentença completa para descrever corretamente a venda da cerveja de malte? A sentença completa, cuja expressão apresenta todos os metadados que compõem a descrição do produto e todos os metadados da venda do produto é descrita na seguinte forma:

"Termo nome do produto + termo nome da marca + termo embalagem (termo do volume + termo tipo da embalagem) + termo pacote + termo NCM + número do item + quantidade do item + valor do item".

A sentença completa é obtida por meio de uma consulta específica no Protégé, com os seguintes requisitos (*DLQuery*):

"NFCeItem and (temCerveja some) and (temPacote some) and (ReferenciaEmbalagem some)".

Como exemplo, foram geradas consultas para cerveja Bohemia:

"NFCeItem and (temCerveja value bohem) and (temPacote value 12x) and (ReferenciaEmbalagem value Lata_350_ml)".

O termo "bohem" foi extraído da lista de termos que o Apriori apresentou ao final da extração de dados, na forma de radical da palavra Bohemia. O resultado da consulta apresentou 11 NFC-*e* com sentenças completas. As 5 primeiras instâncias da classe NFCeItem serão detalhadas neste estudo como resposta a QC01, são elas: *Row* 10561428, *Row* 11546284, *Row* 11747811, *Row* 11747813 e *Row* 13704760 como se apresenta na figura 3. O resultado da sentença completa apresenta os elementos de descrição do produto cerveja Bohemia, o primeiro item é o NCM da cerveja, o segundo item é o nome da marca da cerveja, o terceiro item apresenta o pacote vendido do produto cerveja e o quarto item apresenta a referencia sobre tipo da embalagem e o volume da embalagem. Esses elementos são valores dos *Objects Properties* da classe NFCeItem:

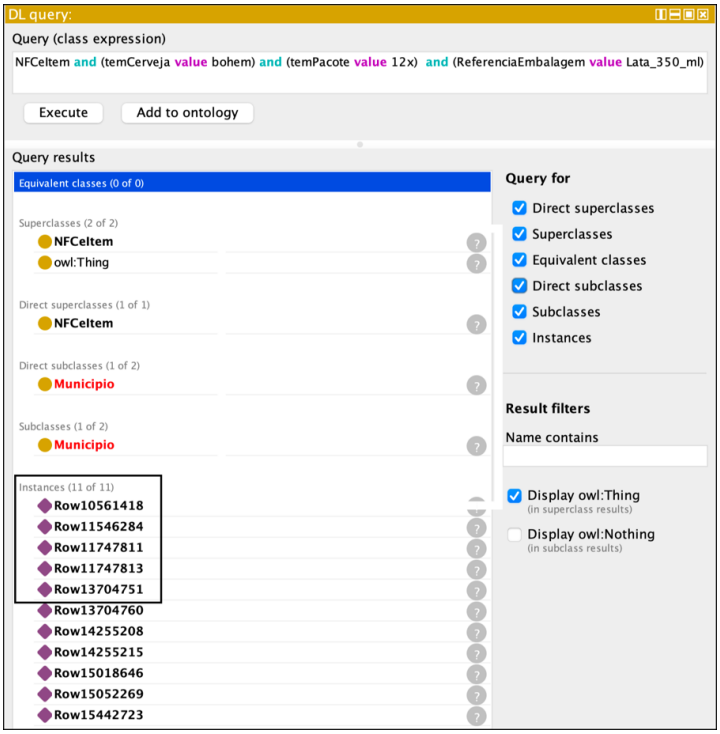

Figura 3: Resultado da DL Query com a seleção de 11 instâncias de sentença completas da NFC-e. Detalhe para as 5 primeiras linhas[3]

Na figura 4, estão os detalhados os valores das 5 primeiras linhas da consulta da cerveja Bohemia:

[3] *Nota.* A DL*Query* busca nome da marca da cerveja Bohemia associado ao pacote com 12 itens associado a embalagem correspondente a lata de 350ml. Fonte: Dados da pesquisa, 2024.

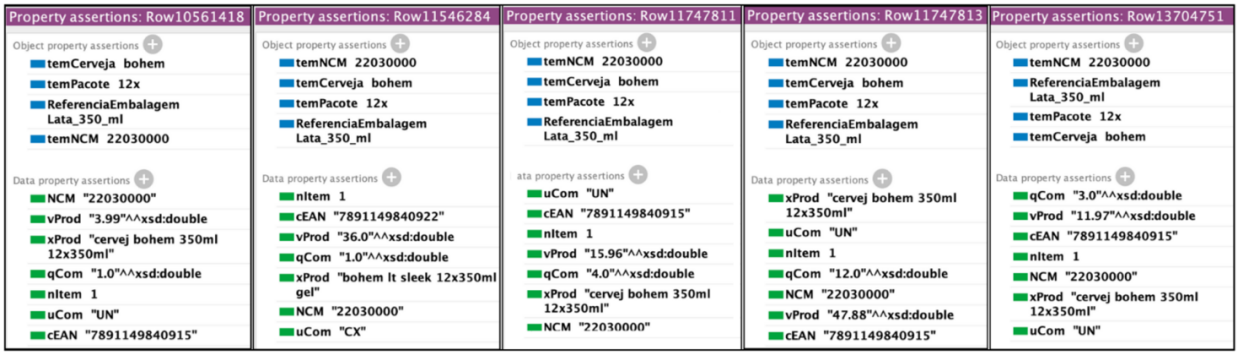

Figura 4: Detalhe das 5 primeiras instâncias da Classe NFCeItem[4]

É possível fazer a busca somente por algum dos metadados da descrição do produto. A figura 5 apresenta uma *DLQuery* que busca identificar a embalagem do produto cerveja Original:

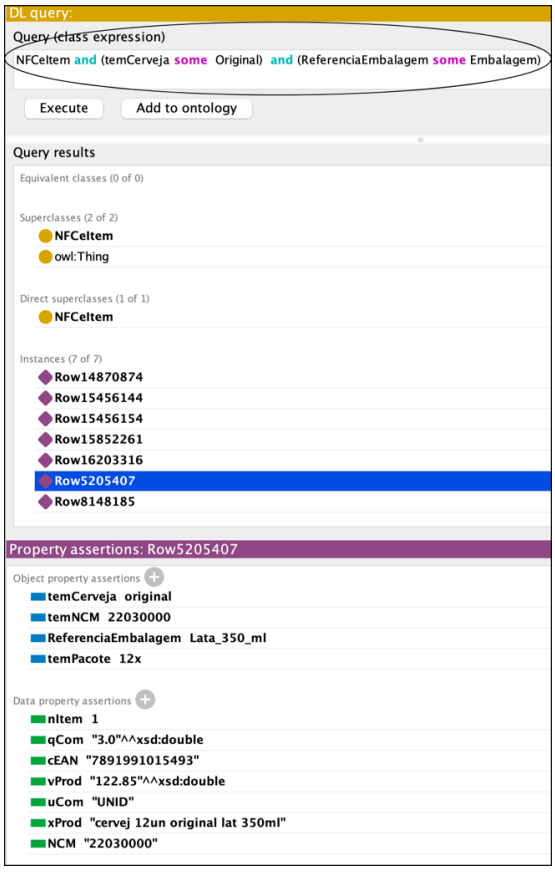

Figura 5: DLQuery "identificação da embalagem do produto cerveja Original".
Nota: Detalhe da linha número 5205407. Fonte: Dados da pesquisa, 2024.

Como é possível observar na figura 5 o resultado da validação da ontologia está detalhado a linha *Row* 5205407 com os dados de interesse da embalagem da cerveja Original e demais descrições do item, caso houver na NFC-*e*, no exemplo acima além da embalagem (com tipo da embalagem e volume) a consulta retornou com o NCM da cerveja Original e com o pacote vendido.

As mesmas DL *Queries* das Figuras 3, 4 e 5 respondem a QC02, pois o NCM é validado no *Object Property Assertions:* "tem NCM 22030000" com a resposta NCM "22030000". A classe NCM tem

[4] *Nota.* Detalhe, da esquerda para direita, das linhas números 10561418, 11546284, 11747811, 11747813 e 13704751. Fonte: Dados da pesquisa, 2024.

instâncias cadastradas com diversos valores como 22030000, 220300222, 22030099, 22030300, 23031000, 22032100, 22033000, 22039000, todos extraídos das análises da base de dados quando sumarizados por NCM com 8 dígitos no início do primeiro evento.

Quanto à QC03, é possível identificar qual a quantidade e preço de determinado produto em determinado período? A resposta desta questão de competência passa pela recuperação e pela utilização dos campos do arquivo da nota fiscal eletrônica transformados em *Data Properties* da classe NFCeItem. Eles compõem campos do .xml da nota fiscal e são conhecidos como: **cEAN**: código de barras do produto (código GTIN); **NCM** (código NCM); **nitem** (ordem sequencial dos itens da nota fiscal); **qCom** (quantidade comercial do produto); **uCom** (unidade comercial do produto); **vProd** (valor total do item da nota); **xProd** (descrição do produto livre). A Figura 6 apresenta os resultados da QC03:

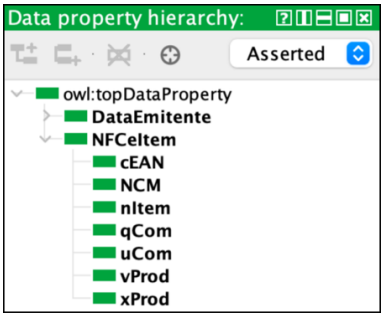

Figura 6: Data Properties da Classe NFCeItem[5]

Na sequência, a QC04, cujo questionamento trata se é possível identificar quais descrições de produto são significativas ou não significativas para um determinado uso, a Figura 7 apresenta o detalhe da linha *Row*1091976 para o produto cerveja Original:

Figura 7: Detalhamento da linha Row1091976 com as informações sobre o produto e venda do produto[6]

---

[5] *Nota.* Os sete campos de *Data Properties* da classe NFCeItem correspondem as seguintes informações: **cEAN**, código de GTIN muitas vezes utilizados como código sequencial de barras do produto, **NCM**, Nomenclatura Comum do Mercosul, **nItem**, número do item dentro do documento fiscal, **qCom**, quantidade comercial, **uCom**, unidade comercial, **vProd**, valor do produto e **xProd**, descrição do produto. Fonte: Dados da pesquisa, 2024.

[6] *Nota.* Resultado da pesquisa com descrição do produto e da venda com informações complementares do GTIN e NCM, metadados logicamente estruturados que permitem a validação do produto em uma segunda camada. Fonte: Dados da pesquisa, 2024.

A QC04 possui uma avaliação subjetiva e a princípio dependeria da avaliação do especialista para classificar como significativas ou não para fiscalização e para determinado uso. Porém, quando se organiza uma sentença completa da venda e da descrição do produto pode se presumir que é significativamente importante para fiscalização quando comparado ao conteúdo disposto de forma desorganizada na base de dados da NFC-*e*. A extração na forma de metadados simplifica a busca e permite comparar termo a termo com a descrição original da nota fiscal que aparece no item: **xProd.** Caso seja do interesse da fiscalização a avaliação de determinada termo ou característica de uma cerveja isso será possível já que a consulta retorna na forma de termos concatenados, caso contrário a as consultas retornarão somente na forma de sentenças completas, como visto nas figuras 2 e 3.

A forma de apresentação do dado como período, NCM, nome da marca da cerveja, determinado volume ou tipo de embalagem também dependerá do critério de aplicabilidade do fiscal. Pode-se entender que os critérios de consulta e recuperação da informação formarão *clusteres* com resultados apropriados para cada critério, oferecendo valor à informação comunicada ao fiscal.

Diferente dos trabalhos avaliados anteriormente para construção de ontologias utilizando PLN e algoritmos não supervisionados, é observado sempre o especialista presente para elaborar a lista de termos candidatos à construção ou enriquecimento. As técnicas de PLN são utilizadas para limpeza dos dados e/ou para obter o significado dos termos por meio de mineração em Corpus de interesse do domínio, ou até mesmo para minerar um tipo sintático pré-determinado como verbos, por exemplo, apenas na fase de enriquecimento da ontologia.

O Apriori é somente utilizado para as regras de associação depois da etapa de conceituação e formalização, oferecendo ao especialista um conjunto de regras para avaliar o resultado das inferências produzidas. No caso das ontologias que tratam do domínio de bebida, especialmente 'cerveja', elas oferecem a descrição do tipo da cerveja, mas não contemplam o nome da marca cerveja nem por classes ou instâncias, inviabilizando a criação de uma sentença completa para descrever o produto vendido, visto que o atributo 'nome da marca da cerveja' é muito utilizado na transação de venda nas NFC-e substituindo, inclusive, o termo principal cerveja', como detalhado em [4].

Apenas um trabalho investigado até o momento uniu conceitos de comercialização do produto e descrição do produto, mas utilizou o conceito genérico de 'item' sem aprofundar nos atributos do item para descrevê-lo.

6. Considerações finais

Com o estudo finalizado, foi possível perceber que, na ontologia 2203NFC-*e,* a aplicação do módulo PLN em conjunto com algoritmo Apriori nas etapas de levantamento e análise deu agilidade na conceitualização e elaboração da lista de termos relevantes, incluindo a identificação de um termo principal 'cerveja' ao qual o NCM 2203.xxxx faz explícita referência, sem a consulta aos especialistas da SEFAZ/AM. A aplicação do módulo de IA nessas fases diferencia este trabalho dos demais pesquisados e permite estendê-lo a qualquer outro NCM pois, independente da descrição do NCM a metodologia de desenvolvimento com o modulo de IA busca um termo principal por importância e número de repetições na base de dados e a partir deste termo principal a associação outros termos como atributos e características para construir a sentença completa.

Com a obtenção dos 210 termos, a partir dos radicais dos produtos descritos nas transações o Apriori, criaram-se as regras de relacionamento entre os termos, avaliadas por meio do Suporte mínimo, *Lift* e Confiança e que garantiram que as associações formadas fossem dependentes dos termos mais frequentes e mais relevantes. Na sequência, os termos e as regras selecionadas pelo Apriori foram formalizados na linguagem OWL; outros conceitos relacionados à comercialização da NFC-*e* foram introduzidos a partir dos dados parametrizados da própria nota fiscal, resultando em um pacote de vendas para qualquer produto comercializado no modelo NF-*e* ou NFC-*e* não apenas de NCM 2203.xxxx.

Os resultados das QCs são relevantes e devem ser considerados, pois representam a realidade dinâmica do setor de fiscalização e do enorme número de NCMs existentes que inviabilizariam a mobilização do especialista para a construção de uma ontologia para cada NCM de interesse. Além

disso, precisam de ferramentas para 'ler' e 'interpretar' cada transação da NF-*e* e NFC-*e*, identificando discrepâncias de descrição e valores de produtos, condição que a ontologia oferece e valida por meio das QCs. outras validações de interesse da fiscalização, mais específicas, podem ser construídas uma vez que a ontologia está construída para suportar a investigação característica do setor e tarefa de fiscalizar e auditar.

A utilização das descrições de cerveja contidas na Resolução n.º 0028/2023 SEFAZ/AM da SEFAZ/AM bem como a utilização das descrições de produtos encontrados nas linhas transacionadas neste estudo enriqueceram os dados das classes que poderiam ser encontrados durante as consultas realizadas pelas *Queries*.

De tal forma, a ontologia serviu para garantir quais informações da nota estão corretas, uma comparação por exemplo, com consultas realizadas diretamente na base com SQL, que só conseguiriam devolver parcialmente as informações porque não possuem inteligência para garantir os relacionamentos das propriedades das classes pesquisadas.

A limitação do trabalho está no tamanho da base pesquisada, uma amostra de 4 meses. A escolha do NCM não se trata de limitação pois a fiscalização trabalha no modelo de auditoria por emitente e por produto de interesse, ou seja, por NCM. Entretanto, os resultados alcançados devem ser espelhados, em um trabalho futuro, para uma base de teste da SEFAZ/AM para ver o comportamento da seleção de termos, se há variação de termos novos, não identificados neste estudo, relevantes os suficientes em importância e número de ocorrência que possam divergir na forma de construção da ontologia e resultado da QC.

Como trabalho futuro desta pesquisa está a produção de repositório NFC-*e* com sentenças completas úteis à fiscalização, ou semicompletas, a depender da necessidade de construção do PMPF ou investigação de preços etc. Outra necessidade é a construção de modelo relacional logico e físico a partir das descobertas do modulo de IA e das relações da ontologia para viabilizar a recuperação da informação nas bases da SEFAZ/AM, um estudo importante a partir, primeiramente, da investigação de vários NCMs de interesse, e deles um modelo relacional generalizado capaz de importar os termos e relacionamentos de forma inteligente.

A ontologia 2203NFC-*e* tem como principal resultado de sua aplicação a reprodução do produto cerveja por sentença completa, com descrição do nome da cerveja, tipo da embalagem, volume, pacote, quantidade e valor individual extraída de qualquer tipo de descrição desordenada do produto no campo "descrição do produto", avaliando quais são completas e quais são úteis em termos de garimpo de uma ou mais de uma propriedade do termo principal. Isso permitiu um avanço na leitura e na interpretação dos dados de venda de cerveja para a fiscalização.

References

[1] Frossard, D. ICMS Genérico. Rio de Janeiro, Editora Ferreira, 2011.

[2] Chung, K.; Yoo, H.; Choe, D. Ambient context-based modeling for health risk assessment using deep neural network. Journal of Ambient Intelligence and Humanized Computing, v. 11, 2020.

[3] Gorayeb, D. M. C.; Gottschalg-Duque, C. Planejamento de um ambiente informacional automatizado para a extração de termos relevantes à fiscalização em nota fiscal eletrônica e a nota fiscal de consumidor eletrônica. XXII Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação, Porto Alegre (RS), 2022.

[4] Gorayeb, D. M. C.; Gottschalg-Duque, C. Proposta de metadados para descrição de produtos da Nota Fiscal de Consumidor Eletrônica (NFC-e) usando Apriori. 2024.(submetido à publicação).

[5] Gruber, T. R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal Human-Computer Studies, v. 43, Padova, Italy, 1995, p. 907-928.

[6] Haridy, S.; Ismail, R.M.; Badr, N.; Hashem, M. An Ontology Development Methodology Based on Ontology-Driven Conceptual Modeling and Natural Language Processing: Tourism Case Study. Big Data and Cognitive Computing 7, no. 2:101, 2023.

[7] Michaelis Dicionário Brasileiro da Língua Portuguesa, Ed. Melhoramentos, 2015. URL: https://michaelis.uol.com.br/moderno-portugues/busca/portugues-brasileiro/cerveja/. Acesso em 26/05/2024.
[8] Resolução n°. 0028 SEFAZ/AM, 2023. URL: https://online.sefaz.am.gov.br/silt/Normas/Legisla%C3%A7%C3%A3o%20Estadual/Resolu%C3%A7%C3%A3o%20GSEFAZ/Ano%202023/Arquivo/RG%200028_23.htm. Acesso em 26/07/2024.
[9] Schulze, M. et al. P2P-O: A Purchase-To-Pay Ontology for Enabling Semantic Invoices. In: ESWC 2021: The Semantic Web. Lecture Notes in Computer Science (), vol 12731. pp 647–663 Springer, Cham, 2021.
[10] Standaert, L.; Yaroslaski, A.; Castro, M. de. Beer Advisor – A beer ontology. Association for the Advancement to Artificial Intelligence, Vancouver, 2021.
[11] Warren, R. (2024). The Beer Ontology. https://rdf.ag/o/beer-en.html. Acesso em 04/03/2024.
[12] Yang, B. Construction of logistics financial security risk ontology model based on risk association and machine learning. Safety Science, v. 123, 2020.
[13] Ontologia 2203NFC-*e*, (2024). URL: https://webprotege.stanford.edu/#projects/6b03286c-6feb-4720-bdf4-692e233fbf54/sharing.
[14] Gil, A. C. Como elaborar projetos de pesquisa. 4. ed. São Paulo: Atlas S.A, 2002.
[15] Marconi, M. de A.; Lakatos, E. M. Fundamentos de metodologia científica. 5. ed. São Paulo: Atlas S.A, 2003.
[16] Flick, U. Métodos de Pesquisa - Introdução à Pesquisa Qualitativa. Tradução de Joice Elias Costa. 3. ed. Porto Alegre: Artmed, 2009.